

Índice

[**Introdução 3**](#_1zdt9aa3n0gw)

[**Tratamento de Dados no Excel 3**](#_g7jkqb16u9a8)

[**Jamovi 5**](#_to7fjuc6q537)

[**R 8**](#_qe5q3cgdlr60)

[**Conclusão 11**](#_wu5hguuj9pmk)

## Introdução

Este relatório, elaborado no âmbito da unidade curricular Análise Exploratória de Dados, tem como intuito preparar, organizar e elaborar uma análise descritiva do conjunto de dados fornecido, em função do contexto e dos objetivos definidos. Com recurso ao Excel, Jamovi e R, temos como alvo a assimilação de todo o conhecimento adquirido em aula, de uma forma mais prática.

Em 2022, após o período de pandemia Covid-19, foi promovido pelo Ministério de Educação o estudo, “Saúde Psicológica e Bem-estar”, das crianças e adolescentes em idade escolar (do pré-escolar ao 12º ano), em diferentes regiões do norte de Portugal. Todos os dados em análise provêm de um questionário realizado a 996 indivíduos, entre 11 e 20 anos, relativamente a perguntas enquadradas nos objetivos definidos e informações de caracterização das crianças e jovens.

Numa fase inicial, é feito um tratamento prévio de dados com ferramentas técnicas para uma melhor organização das amostras a serem contabilizadas. Posto isto, procedemos à montagem de gráficos e tabelas para análise, compreensão e comparação dos resultados obtidos no questionário.

## Tratamento de Dados no Excel

Variáveis

Para uma organização de dados mais eficiente, alterou-se primeiramente os códigos originais representativos das variáveis para nomes mais indicadores das mesmas, por exemplo, substituindo os nomes “P3”, “P5”, “P8\_rec”, “v41\_depressao”, “v46\_depressao”, “v49\_depressao”, “v52\_depressao”, “v53\_depressao”, “v57\_depressao”, “Sono\_Q556” e “Ecra\_Q558” para “Idade”, “Nível Escolar”, “Satisfação”, “Dificuldade em tomar iniciativa”, “Expectativa de futuro”, “Depressão”, “Pouco entusiasmo”, “Valor pessoal”, “Pensamentos suicidas”, “Horas de sono” e “Tempo de ecrã”, respetivamente.

De seguida, selecionou-se no guia Ver > Congelar Painéis > Congelar Primeira Linha, para tornar sempre visível o tipo da informação a ser analisada. Da mesma forma, para podermos visualizar mais rapidamente as respostas dos inquiridos, acrescentou-se novas colunas, correspondentes aos valores qualitativos, à direita das colunas originais com o respetivo código, recorrendo às funções vlookup e hlookup.

Valores Omissos

Para encontrar os valores omissos existentes na base de dados, selecionou-se a opção “BLANK”, através da opção “Filtro” posto isto, recorremos à imputação do valor 99, introduzindo manualmente este valor nas células vazias.

Porém também recorremos a funções com esta finalidade, tal como: IF(OR(ISBLANK());99;VLOOKUP()).

A falta de dados nas variáveis é uma situação muito frequente que traz implicações na análise e interpretação dos dados.

Correção de Erros

Para evitar distorções nos resultados ou chegar a conclusões erróneas, utilizou-se o comando “Filtro”, para detetar respostas equivocadas, ou fora do intervalo definido, como por exemplo, o valor “188” na variável “Distrito”, o “33” na variável “Ciclo escolar” e o “12” na variável “Pouco entusiasmo”. Nestes casos, assume-se que o indivíduo se enganou ao preencher o questionário. De modo a apenas trabalhar-se com dados fiáveis, retirou-se estas amostras, deixando as células errôneas em branco.

Validação de variáveis

Como pedido no enunciado, escolhemos como variável quantitativa o género codificado, coluna E, e como variável qualitativa o género em si, sem ser codificado, coluna F. Utilizámos como regra de validação para a coluna E, o intervalo de resposta [1, 4]. Para a coluna F, como critério colou-se as 4 opções de resposta, “Masculino”, “Feminino”, “Outro”, “Prefiro não responder”. Desta forma, todos os valores fora deste intervalo fechado (coluna E) e fora destas opções (coluna F), desencadeiam um aviso, de modo a atrair atenção para os mesmos, facilitando a sua correção.

Tabelas

A principal diferença entre uma tabela de frequência absoluta e uma tabela de frequência dinâmica é que a primeira é estática e mostra a contagem de ocorrências de cada valor em um conjunto de dados, enquanto a segunda é atualizada em tempo real à medida que novos dados são adicionados.

Na tabela de frequência absoluta a % em linha corresponde à percentagem de cada valor nessa mesma linha nas colunas que se seguem, ou seja, a percentagem de pensamentos suicidas por gênero. Utilizamos também filtros na tabela para ocultar não respostas.

Na tabela de frequência absoluta, comparamos o nível de pensamentos suicidas com o gênero do aluno e observamos um grande número (54%) de alunos que identificaram o seu gênero como “Outro”, têm um índice preocupante de pensamentos suicidas. O segundo maior número, com uma diferença bastante grande face aos restantes, de (19%) dos estudantes que preferiram não responder ao parâmetro do seu gênero. Também é possível observar que (80%) dos alunos do sexo masculino apresentam menores tendências para pensamentos suicídas.

Na tabela de frequência dinâmica, escolhemos contar o número de estudantes com diferentes graus de pensamentos suicidas, a maioria não apresenta estes pensamentos intrusivos, sendo este número de 690 num total de 959 alunos. No outro lado do espectro, temos 49 alunos com severos pensamentos deste gẽnero.

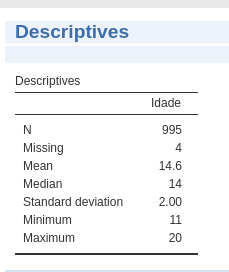
Na tabela de cruzamento, analisamos os níveis de satisfação e depressão em função do ciclo escolar em que estão.

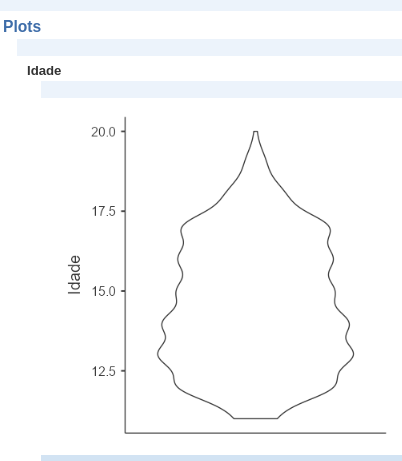
Os alunos do ensino secundário e do básico apresentam valores ligeiramente diferentes, mas não significativos, não se podendo então tirar conclusões sobre esta comparação. No entanto, é possível concluir que a maioria dos alunos (418), não se sente deprimido e tem uma média alta de satisfação de nível 8 na escala. Sendo, por outro lado, o número de alunos com um grande índice depressivo de apenas 37 num total de 949.

## Jamovi

Através do jamovi, elaboramos vários gráficos e 3 tabelas com medidas descritivas.

Primeiramente, analisamos a variação das idades dos 995 alunos, com apenas 4 não respostas neste parâmetro. Concluindo que, a média é de 14 anos, assim como a mediana. Existindo uma frequẽncia maior de alunos dos 12 aos 17 anos, sendo as idades mínimas e máximas 11 e 20 respetivamente. Para além disto, o desvio padrão é de 2, o que significa que há uma dispersão relativamente alta entre os valores e uma amplitude de 9.

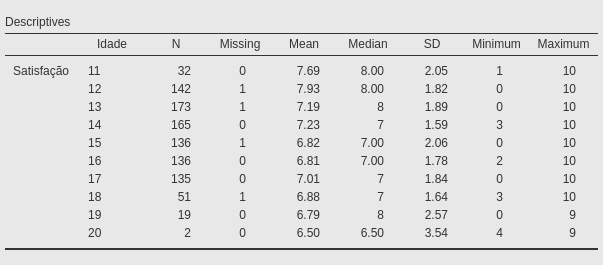


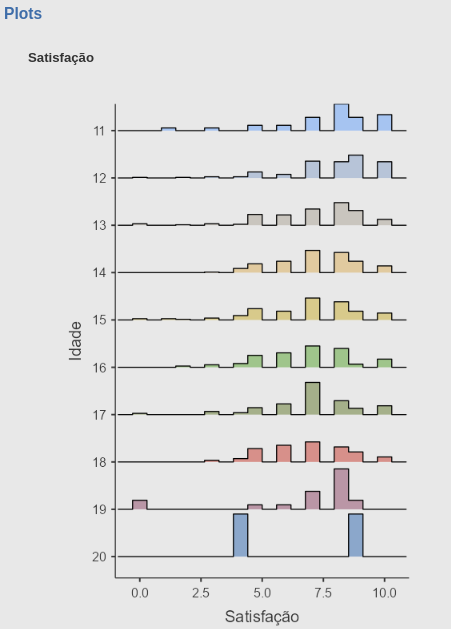


Na próxima tabela, analisamos o tempo de ecrã, em horas, e as horas de sono, pela idade dos alunos. Concluindo-se que, relativamente ao tempo gasto em ecrãs, a média ronda as 5 horas, com um mínimo e um máximo de 1 e 10 horas, respetivamente.

Analisando as horas de sono, apresenta-se como mínimo 4 horas, e um máximo de 10, com uma média de 8.

Seguidamente, numa próxima tabela de medidas descritivas, analisamos a satisfação dos alunos, por idade. Obtendo uma média e uma mediana a rondar os 7, numa escala de 0 a 10. O desvio padrão é de 2, concluindo-se que os valores são relativamente dispersos uns dos outros, em certas idades, aos 13 e 18, o mínimo satisfatório é de 3, e de 2 aos 16 anos. Obtemos valores inferiores nas restantes idades, de 0 e 1 como valores mínimos, e máximos de 10 para todas as idades, exceto aos 20 e 19 anos, que é de 9, não sendo muito distinto.





Uma vez que, obtivemos apenas 2 casos de alunos com 20 anos, não consideramos tanto estes dados na nossa análise. Concluímos também que, para as variáveis que escolhemos analisar, os dados são semelhantes para qualquer uma das idades.

## R

Para uma melhor análise e compreensão dos dados, utilizamos as seguintes bibliotecas do R. Para abrir um ficheiro do tipo xlsx, formato de excel, utilizámos a biblioteca openxlsx, a tidyverse para , a ggplot2 para observação de gráficos variados, a jmv para análise de dados multivariados e modelagem estatística, a treemap para visualização de dados em forma de mapas de árvore. Permitindo assim a visualização da distribuição dos dados em diferentes grupos e subgrupos de forma hierárquica. A biblioteca descr para cálculo de estatísticas descritivas, como média, desvio padrão, mediana, quartis, entre outros. A moments para cálculo de momentos estatísticos, como média, variância, assimetria e curtose. Sendo assim útil para analisar a distribuição de dados e detectar desvios da normalidade. A flextable para a organização dos dados em tabelas, sendo possível personalizar a aparência da tabela, incluindo a formatação da fonte, as cores, bordas e muito mais. A olsr para análise de regressão linear. E, finalmente a foreign para a leitura e escrita de arquivos de dados em outros formatos, como SPSS, SAS, Stata e outros. Pode ser útil para trabalhar com dados que foram recolhidos ou armazenados noutro software estatístico, no nosso caso, jamovi e excel.

Primeiramente, importamos o ficheiro de excel, utilizando a função read.xlsx() e mudámos o nome de duas variáveis, ID para CC e NUTII para Região, usando a função names().

Valores omissos

Como já alterámos previamente no Excel os valores omissos para o número 99, agora inserimos NA da seguinte forma: dt$Horas.de.sono[dt$Horas.de.sono == 99] <- NA, a fim de podermos utilizar as seguintes funções para identificar os valores omissos da variável "Horas de sono": unique(dt$Horas.de.sono) para ver as respostas utilizadas e is.na(dt$Horas.de.sono), que retorna uma matriz lógica, indicando se a resposta do aluno foi nula ou não.

De modo a imputar os valores omissos da variável escolhida, com a mediana dos seus valores recorremos às seguintes funções, round(median(dt$Horas.de.sono, na.rm=TRUE)), que retorna a mediana arredondada da nossa variável escolhida, que neste caso é 8 e, finalmente, replace(dt$Horas.de.sono,is.na(dt$Horas.de.sono), 8) que verifica os valores omissos e imputa o valor da mediana (8).

Tabelas de frequências com valores absolutos e percentagens

De modo a realizar tabelas de frequência com valores absolutos e percentagens, recorremos às seguintes funções, table(cd$Horas.de.sono), esta função é utilizada para criar uma tabela de frequência a partir de um vetor. Aqui, a tabela é criada a partir da variável "Horas de sono" que está presente no data frame "cd". A função conta o número de ocorrências de cada valor único na variável e apresenta esses valores em forma de tabela.

Posto isto, recorremos à seguinte função, round(prop.table(thorassono)\*100,1), a função prop.table() é usada para calcular a proporção de cada valor único na tabela de frequência. O resultado é uma matriz de proporções que somam 1.0. A função round() é usada para arredondar essas proporções para um número específico de casas decimais. As proporções são arredondadas para uma casa decimal. Em seguida, o resultado é multiplicado por 100 para obter as percentagens.

Depois disto, combinamos estas duas tabelas usando a seguinte função, bind\_rows(thorassono, thorasperc). Desta forma, as duas tabelas são combinadas em um único data frame com duas colunas: "Freq" (contagem) e "Percent" (percentagens). Esta combinação permite comparar a contagem e as percentagens de cada valor único na variável "Horas de sono" em uma única tabela.

Medidas descritivas de variáveis quantitativas

Em termos de medidas descritivas, fizemos uma tabela de medidas descritivas com as horas de sono dos alunos, concluindo que a média é de 7,6 horas, a mediana 8, o desvio padrão de 1.2 o que significa que os valores estão um pouco dispersos sendo então a variância de 1.4, uma vez que os valores são um pouco diferentes entre si também. Os valores mínimo e máximo são de 4 e 10, respetivamente. A assimetria é de -0,5 significando que a média é menor ou igual à mediana e menor ou igual à moda. A curtose é de 3.5 logo, apresenta uma função de distribuição leptocúrtica. Concluímos também que, 25% dos alunos dorme menos de 7 horas, e 75% dos alunos dorme menos de 9 horas.

Gráficos

Em termos de gráficos, fizemos um pie chart com informação sobre a quantidade de horas que os alunos dormia, com esta finalidade agrupamos os dados em 3 grupos, menos de seis horas, de 7 a 9 horas e mais de 10 horas de sono. Concluindo que a maioria dos alunos dorme entre 7 a 9 horas.

Seguidamente, fizemos um gráfico do tipo boxplot, sobre a satisfação dos alunos com a sua vida. A partir deste gráfico analisamos que a mediana (2º quartil) é de 7. Para além disto, conseguimos afirmar que 75% dos alunos medem a sua satisfação com um nível inferior a 8 e 25% inferior a 6, sendo que os valores máximo e mínimo são de 10 e 0, respetivamente.

R Markdown

Seguidamente recorremos ao R Markdown para construir um relatório com a interpretação dos resultados, estruturados de acordo com os objetivos do estudo.

## Conclusão

Com base nos dados recolhidos e analisados neste relatório de ciência de dados sobre a saúde mental dos estudantes, podemos concluir que a saúde mental dos estudantes é uma questão crítica que merece mais atenção e recursos. Os resultados mostram que um número significativo de estudantes enfrenta problemas de saúde mental, como questionamentos sobre o seu próprio valor pessoal, satisfação com a vida, suscetibilidade à depressão e, em certos casos, pensamentos suicidas. Estes problemas podem afetar negativamente o desempenho académico e o bem-estar geral dos estudantes. É essencial que as instituições educacionais, bem como as autoridades governamentais, tomem medidas para oferecer apoio adequado à saúde mental dos estudantes, incluindo a disponibilização de recursos de saúde mental, a implementação de programas de bem-estar estudantil e a consciencialização sobre a importância do equilíbrio mental. Este relatório destaca a importância de continuar a recolher e analisar dados sobre a saúde mental dos estudantes para que possamos compreender melhor esta questão e trabalhar para melhorar a saúde mental dos estudantes em todo o mundo.